DOCUMENT RESUME

ED 067 401 TM 001 796

AUTHOR	Sween, Joyce; Campbell, Donald T.
TITLE	A Study of the Effect of Proximally Autocorrelated Error on Tests of Significance for the Interrupted Time Series Quasi-Experimental Design.
INSTITUTION	Northwestern Univ., Evanston, Ill.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C. Educational Media Branch.
REPORT NO	Proj-C998
PUB DATE	Aug 65
CONTRACT	OEC-3-20-001
NOTE	43p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Correlation; *Data Analysis; *Mathematical Models; *Measurement Techniques; Statistics; Tests; *Tests of Significance
IDENTIFIERS	Double Extrapolation Technique; Mood Test; Walker Lev Test 3

ABSTRACT

The primary purpose of the present study was to investigate the appropriateness of several tests of significance for use with interrupted time series data. The second purpose was to determine what effect the violation of the assumption of uncorrelated error would have on the three tests of significance. The three tests were the Mood test, Walker-Lev Test 3, and Double Extrapolation Technique. The procedure was basically that of generating a large number of time series having specified characteristics and performing the tests of significance on each generated time series. The results of the study indicated that the three tests of significance are appropriate for use on data of interrupted time series form. Tables and figures illustrate the text. (Author/DB)

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED 067401

A STUDY OF THE EFFECT OF PROXIMALLY AUTOCORRELATED ERROR ON TESTS OF SIGNIFICANCE FOR THE INTERRUPTED TIME SERIES QUASI-EXPERIMENTAL DESIGN*

Joyce Sween and Donald T. Campbell

August 1965

Northwestern university

TM 001 796

*Supported by Project C998, Contract 3-20-001, Educational Media Branch, Office of Education, U.S. Department of Health, Education, and Welfare, under provisions of Title VII of the National Defense Education Act.

FILMED FROM BEST AVAILABLE COPY

A Study of the Effect of Proximally Autocorrelated Error on Tests of Significance for the Interrupted Time-Series Quasi-Experimental Design

Joyce Sween and Donald T. Campbell

Northwestern University

The time series experiment has long been a common research design in the biological and physical sciences. However, with psychological and sociological data certain problems of analysis and interpretation occur when the interrupted time series design is employed (Campbell, 1963; Campbell & Stanley, 1962; Holtzman, 1963).

One of the problems which is of particular concern to the social scientist (for whom the magnitude and clarity of effects is not always as clear-cut as in the biological and physical sciences) is testing for the significance of the change, X. It is desired to have some statistical test of significance that would distinguish the effect of an intervening event or experimental variable from purely random fluctuation. Although there are no tests of significance that are completely suitable to the time series situation, several possibilities for such a test do exist and merit consideration (Campbell, 1963; Campbell & Stanley, 1962).

Another problem of fundamental concern to the social scientist is the possibility of sequential dependency in successive observations of the time series. Measurements made at time points which are closer together may be more strongly related than measurements made farther apart in time. Statistical models on which the tests of significance are based generally assume that the observed values exhibit independent error. Thus, the significant

autocorrelation which exists when a given observation is dependent upon the observations preceding it violates the assumption of independence in these tests. Although this circumstance does not prevent estimation (by least square methods) of the regression coefficients which the various tests of significance use, the classical formulas for the mean errors of the estimated coefficients are in actuality no longer valid. When successive observations through time are correlated, perhaps, to a lag of several points, use of existing tests of significance may be inappropriate.

The primary purpose of the present study was to investigate the appropriateness of several tests of significance for use with interrupted time series data. The three main tests were those described by Mood (1950, pp. 297-298; pp. 350-358) and Walker and Lev (1953, pp. 390-395; pp. 399-400).

The interrupted time series experiment as considered in the present study consists of a finite sequence of observations which are real-valued measurements of some individual or group obtained at n successive equally spaced points in time. An experimental change is then introduced or a major change of conditions occurs at some point in the time series. In the present study, the experimental change or treatment is assumed to occur midway between two consecutive measurements. It can be diagrammed as follows:

Interrupted Time Series Design

Observation at time t_i:	y_1	y_2	...	y_m	y_{m+1}	y_{m+2}	...	y_n
Time:	t_1	t_2	...	t_m ↑	t_{m+1}	t_{m+2}	...	t_n
				experimental change				

The y_i are the measurements recorded as time t_i. The values $y_i (i = 1,2, \ldots m)$ obtained prior to the treatment are referred to as pre-change values. The

values y_i ($i = m+1, m+2, \ldots n$) obtained after the treatment are post-change values. If the treatment has produced an effect, a discontinuity of measurements is recorded in the time series. The statistical tests of significance which were investigated in the present study as possible techniques for distinguishing such a treatment effect from purely random fluctuation are described in detail below:

(a) Mood Test. This is a t test (Mood, 1950, pp. 297-298) for the significance of the first post-change observation from a value predicted by a linear fit of the pre-change observations. The formula for t is

$$t = \frac{(Y_0 - \hat{Y}_0)/\sigma_u}{\sqrt{m\hat{\beta}^2/(m-2)\sigma^2}}$$

The estimate, $\hat{Y}_0$, of the first post-change value is obtained from the pre-change regression estimates

$$\hat{\beta} = \frac{\sum_{i=1}^{m} t_i Y_i - (\sum_{i=1}^{m} t_i)(\sum_{i=1}^{m} Y_i)/m}{\sum_{i=1}^{m} t_i^2 - (\sum_{i=1}^{m} t_i)^2/m}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{t}$$

where m = number pre-change points
$i = 1, 2, \ldots m$
$\bar{Y} = \sum Y_i/m$
$\bar{t} = \sum_{i=1}^{m} t_i/m$

The difference between the estimated value and the obtained first post-change value, Y_0, is used in the t test. Since $\sigma_u = \sqrt{\sigma^2\left[\frac{m+1}{m} + \frac{(t_0-\bar{t})^2}{\sum_{i=1}^{m}(t_i-\bar{t})^2}\right]}$ and an estimate, $\hat{\sigma}^2$, of σ^2 is given by $\frac{1}{m}\sum_{i=1}^{m}(Y_i - \hat{\alpha} - \hat{\beta}t_i)^2$, the computation formula for the t test of the difference between the obtained and predicted post-change point is given by

$$t = \frac{|Y_0 - \hat{Y}_0|}{\sqrt{\left(\frac{m+1}{m} + \frac{(t_0-\bar{t})^2}{\sum_{i=1}^{m}(t_i-\bar{t})^2}\right)\left(\frac{\sum_{i=1}^{m}(Y_i - \hat{\alpha} - \hat{\beta}t_i)^2}{m-2}\right)}}$$

The denominator is the standard error of the difference for t. The df for t is (m-2). The significance of the obtained t is evaluated by reference to a standard t table.

(b) Walker-Lev Tests. Walker and Lev (1953, pp. 390-395) provide the following tests of significance. These tests are useful in determining whether differences exist between the regression equations for pre- and post-change groups.

Test One.

This is a test of the hypothesis of common slope. The F ratio is given by $F = \frac{S_1}{S_2} \cdot \frac{(N-4)}{1}$

where S_1 = sum of squares of the common within groups regression estimates from the separate group regression estimates

$$\sum_{i=1}^{2} \sum_{j=1}^{N_i} \left[(\hat{\alpha}_i + \hat{\beta}_i t_{ij}) - (\hat{\alpha}_{\omega_i} + \hat{\beta}_\omega t_{ij}) \right]^2$$

S_2 = sum of squares of the obtained occasion values from the separate group regression estimates

$$\sum_{i=1}^{2} \sum_{j=1}^{N_i} \left[Y_{ij} - (\hat{\alpha}_i + \hat{\beta}_i t_{ij}) \right]^2$$

N = total number of occasions in time series
i=1,2 [1=prechange group; 2=postchange group]
$j = 1, 2, \ldots N_i$

Formulas for the common within groups slope and intercept are

$$\hat{\beta}_\omega = \frac{CTY_\omega}{CTT_\omega}$$

$$\hat{\alpha}_{\omega_i} = \bar{Y}_i - \hat{\beta}_\omega \bar{t}_i$$

Formulas for the separate group slopes and intercepts are

$$\hat{\beta}_i = \sum_{j=1}^{N_i} t_{ij} Y_{ij} - \left(\sum_{j=1}^{N_i} t_{ij} \right) \left(\sum_{j=1}^{N_i} Y_{ij} \right) / N_i$$

$$\hat{\alpha}_i = \bar{Y}_i - \hat{\beta}_i \bar{t}_i$$

The computational formulae for the numerator and denominator sum of squares for the F ratio are

$$S_1 = \sum_{i=1}^{2} \frac{CTY_i}{CTT_i} - \frac{C^2TY_\omega}{CTT_\omega}$$

$$S_2 = CYY_\omega - \sum_{i=1}^{2} \frac{C^2 TY_i}{CTT_i}$$

where

$$CTY_i = \sum_{j=1}^{N_i} t_{ij}\, Y_{ij} - \left[\frac{\sum_{j=1}^{N_i} t_{ij} \sum_{j=1}^{N_i} Y_{ij}}{N_i} \right]$$

$$CTT_i = \sum_{j=1}^{N_i} t_{ij}^2 - \frac{\left(\sum_{j=1}^{N_i} t_{ij} \right)^2}{N_i}$$

$$CYY_i = \sum_{j=1}^{N_i} Y_{ij}^2 - \frac{\left(\sum_{j=1}^{N_i} Y_{ij} \right)^2}{N_i}$$

$$CTY_\omega = \sum_{i=1}^{2} CTY_i$$

$$CTT_\omega = \sum_{i=1}^{2} CTT_i$$

$$CYY_\omega = \sum_{i=1}^{2} CYY_i$$

The resulting F ratio is evaluated with 1 and N-4 degrees of freedom.

Test Two.

This is a test of the hypothesis that the slopes of the pre- and post-change groups are equal to zero when it has been established that $\hat{B}_i = \hat{B}_\omega$ (that is, F of Test One is not significant). The variance ratio is given by

$$F = \frac{C^2 TY_\omega}{CTT_\omega \cdot CYY_\omega - C^2 TY_\omega} \cdot \frac{N-4}{1}$$

and is evaluated with 1 and N-4 degrees of freedom.

Test Three.

This is a test of the hypothesis that a single regression line fits both the pre- and post-change groups. The F ratio is given by

$$F = \frac{S_b}{S_\omega} \cdot \frac{N-3}{1}$$

where S_b = sum of squares of common within groups regression estimates from the total series regression estimates

$$\sum_{i=1}^{a}\sum_{j=1}^{N_i}\left[(\hat{\alpha}_{w_i}+\hat{\beta}_w t_{ij})-(\hat{\alpha}_T+\hat{\beta}_T t_{ij})\right]^2$$

S_w = sum of squares of obtained occasion values from the common within groups regression estimates

$$\sum_{i=1}^{a}\sum_{j=1}^{N_i}\left[Y_{ij}-(\hat{\alpha}_{w_i}+\hat{\beta}_w t_{ij})\right]^2$$

Formulas for the total series slope and intercept are

$$\hat{\beta}_T = \frac{CTY_T}{CTT_T}$$

$$\hat{\alpha}_T = \bar{y} - \hat{\beta}_T \bar{t}$$

The computational formulae for the numerator and denominator sum of squares for the F ratio are

$$S_b = CYY_B + \frac{C^2TY_w}{CTT_w} - \frac{C^2TY_T}{CTT_T}$$

$$S_w = S_1 + S_2$$

where

$$CTY_T = \sum_{i=1}^{a}\sum_{j=1}^{N_i} t_{ij}Y_{ij} - \left[\left(\sum_{i=1}^{a}\sum_{j=1}^{N_i} t_{ij}\right)\left(\sum_{i=1}^{a}\sum_{j=1}^{N_i} Y_{ij}\right)\Big/ N_i\right]$$

$$CYY_T = \sum_{i=1}^{a}\sum_{j=1}^{N_i} Y_{ij}^2 - \left[\left(\sum_{i=1}^{a}\sum_{j=1}^{N_i} Y_{ij}\right)^2\Big/ N\right]$$

$$CTT_T = \sum_{i=1}^{a}\sum_{j=1}^{N_i} t_{ij}^2 - \left[\left(\sum_{i=1}^{a}\sum_{j=1}^{N_i} t_{ij}\right)^2\Big/ N\right]$$

$$CYY_B = CYY_T - CYY_w$$

The resulting F ratio is evaluated with 1 and N-3 degrees of freedom.

(c) "Double Extrapolation" Technique. This test is concerned with the significance of the difference between two separate regression estimates of the y-value for time t_o which lies midway between the last pre-change point and the first post-change point. Reference to the Interrupted Time Series Design diagrammed on page 2 will indicate that this point lies midway between t_m and t_{m+1}. Assuming that the points are equally spaced in time, t_o is equal to $t_{m+.5}$.

The first regression estimate for y_o is obtained from the pre-change values; the second regression estimate for y_o, from the post-change values. The formulas for the two regression estimates are

$$\hat{\beta}_i = \frac{\sum_{j=1}^{N_i} t_{ij} y_{ij} - \left(\sum_{j=1}^{N_i} t_{ij}\right)\left(\sum_{j=1}^{N_i} y_{ij}\right)\Big/ N_i}{\sum_{j=1}^{N_i} t_{ij}^2 - \left(\sum_{j=1}^{N_i} t_{ij}\right)^2 / N_i}$$

$$\hat{\alpha}_i = \bar{Y}_{ij} - \hat{\beta}_i \bar{t}_{ij}$$

where $i = 1,2$ [1 = prechange group; 2 = postchange group]
N_i = number occasions in group i
$j = 1,2,\ldots N_i$

Thus, the two estimates for y_o are given by $\hat{Y}_{o_i} = \hat{\alpha}_i + \hat{\beta}_i t_o$. The difference between these two estimates can be evaluated by the t-ratio $t = \frac{|\hat{Y}_{o_1} - \hat{Y}_{o_2}|}{S_D}$ and its significance determined by reference to a t table with $N_1 + N_2 - 4$ degrees of freedom. The standard error of the difference for t depends upon the variability of the pre- and post-change values, N_1 and N_2, the relation between t and y, and the distance of t_o from $\bar{t}_1$ and $\bar{t}_2$. The formula for S_D is given in Walker and Lev (p. 400) as

$$S_D = \sqrt{\frac{\left[\sum_{i=1}^{2}\left(CYY_i - \frac{C^2 TY_i}{CTT_i}\right)\right]\left[\frac{N}{N_1 N_2} + \sum_{i=1}^{2}\frac{(t_o - \bar{t}_i)^2}{CTT_i}\right]}{N_1 + N_2 + 4}}$$

where

$$CY_iY_i = \sum_{j=1}^{N_i} Y_{ij}^2 - \left[\left(\sum_{j=1}^{N_i} Y_{ij}\right)^2 / N_i\right]$$

$$CTT_i = \sum_{j=1}^{N_i} t_{ij}^2 - \left[\left(\sum_{j=1}^{N_i} t_{ij}\right)^2 / N_i\right]$$

$$CTY_i = \sum_{j=1}^{N_i} t_{ij} Y_{ij} - \left[\left(\sum_{j=1}^{N_i} t_{ij} \sum_{j=1}^{N_i} Y_{ij}\right) / N_i\right]$$

The second purpose of the present study was to determine what effect the violation of the assumption of uncorrelated error would have on the three tests of significance. Do the tests become inappropriate when there is positive autocorrelation between points?

The scientist usually does not have knowledge of the degree of sequential dependency which exists in a given time series. The first serial correlation coefficient (of lag one) is used to determine whether the observations of a given time series can be regarded as consisting of independent error only (messures not correlated). The serial correlation coefficient tests the hypothesis that the order of dependence in the time series is zero against the alternative that it is one. Serial correlation coefficients of lags higher than one can also be considered. (If the serial correlation is that of a variable with itself, it may be referred to as an auto-correlation coefficient.)

A serial correlation coefficient of lag one, r_1, is obtained by pairing observations one time unit apart; that is, the first observation is paired with the second, the second with the third, etc., throughout the

entire time series until the last observation is reached. The product-moment correlation is computed using the resulting pairs of observations. In a similar manner, the second serial correlation coefficient, r_2, is obtained by pairing successive observations two time units apart and r_m is obtained by pairing observations m units apart. The basic formula for r_a is given below:

$$r_a = \frac{n \sum_{i=1}^{n} Y_i Y_{i+a} - \sum_{i=1}^{n} Y_i \sum_{i=a}^{n} Y_{i+a}}{\sqrt{\left[n \sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right]\left[n \sum_{i=a}^{n} Y_{i+a}^2 - \left(\sum_{i=a}^{n} Y_{i+a}\right)^2\right]}}$$

where N = total number of occasions in the time series
n = N-a
i = 1,2,...n

The model of autocorrelation used in the present study is basically one of proximity, namely, it was assumed that measurements made closer together in time would be more strongly related than measurements made further apart in time. This means that in instances where a significant autocorrelation r_a exists, increasingly larger autocorrelations should exist for r_{a-1}, r_{a-2}, ... and r_1, respectively.

Since the slope of the line contributes to serial dependency between points, interest in the present study is in the autocorrelations of departures from the line of best fit for the total series. However, because true effects increase proximal autocorrelation, differences from the separate pre- and post-change regression lines give better estimates of existing autocorrelations in instances of true effects. So as not to penalize true effects in the data version of the program (a computer program which performs the tests of significance on input data of interrupted time series form is

presented in Sween and Campbell, 1965), autocorrelation coefficients based on departures from the separate regression lines are also computed. Although only departures from the line of best fit for the total series were used in the present study, in further work autocorrelations based on differences from separately fitted regression lines will also be utilized to increase comparability of the Monte Carlo results with actual experimental data.

METHOD

The procedure was basically that of generating a large number of time series having specified characteristics and performing the tests of significance on each generated time series. In this way, distributions of t's and/or F's for the three tests of significance were obtained. The distributions were then examined to determine how satisfactory each of the three tests of significance were in terms of the risk (α) of rejecting the null hypothesis (the experimental change X had no effect) when it was true.

The time series were constructed so that the hypothesis of no effect was true. Normal random error was added to each "true" point to produce a time series of N observed values such that $y_{i(observed)} = y_{i(true)} + error_i$ for $i = 1,...N$. Two general types of error, independent error and/or correlated error, could be added to the "true" line values. When the null hypothesis is true and the assumptions for use of the tests of significance are met, the theoretical values from the t and F tables should be exceeded by chance only 1% and 5% of the time. Thus, when only independent error is added, the discrepancies between the theoretical values and the obtained percent of significant t's and F's should indicate how suitable the test of significance is in the interrupted time series situation. In addition, the

obtained percentage of t's and F's which exceed the theoretical values when sequential dependency between points is built in should indicate how the violation of the assumption of independence of error may further restrict the usefulness of the tests.

Each generated time series could be varied with respect to (a) the number of pre- and post-change occasion values,(b) the slope of the "true" line, (c) the degree of autocorrelation between points, and (d) the total error variance about the true line values. In the present study the following combinations of pre- and post-change occasion values, "true" line slopes, autocorrelated error, and error variance were used:

(a) Pre- and post- change sample sizes of 10, 20, and 100 were used to yield time series with the following number of total occasion values:

Total Occasions, N = 20 (pre = 10; post = 10)
Total Occasions, N = 40 (pre = 20; post = 20)
Total Occasions, N =200 (pre =100; post =100)

In the presentation of the results these three conditions of sampling are referred to in terms of the total occasions in the series as N = 20, N = 40, and N = 200. The degrees of freedom and critical values for the Mood test of significance are, however, based only upon the pre-change points of 10, 20, and 100. For the Walker-Lev Test 3 and Double Extrapolation Technique the total series points of 20, 40, and 200 are used.

(b) The slope of the true line was specified as 0.0, 1.0, and 20.0.

(c) The degree of autocorrelated error was specified as zero (independent error only), one, two, and three (correlated error for measurements one, two, and three time lags apart).

(d) Total error variance about the true line was specified at 1.00 and 5.00 and normal random error was drawn from Gaussian distributions of zero mean and appropriate standard deviations to yield equal error variance about the true line for all degrees of autocorrelated error.

For the total error variance specified as 1.00, the standard deviations of the normal distributions from which the normal random errors were drawn would be

1.00 (unique error only)
.58 (unique plus error of lag 1)
.50 (unique plus error of lag 2)
.45 (unique plus error of lag 3)

For a total error variance specified as 5.00, the standard deviations of the normal distributions from which the normal random errors were drawn would be

2.24 (unique error only)
1.29 (unique plus error of lag 1)
1.12 (unique plus error of lag 2)
1.00 (unique plus error of lag 3)

One thousand sets of time series were generated for each type of autocorrelated error in various combinations with the options of sample size, true line slope, and total error variance listed in a, b, and d above. The following tests of significance were performed on each of the 1000 sets of generated data: Mood test, Walker-Lev Test 3, Double Extrapolation technique. [The Walker-Lev Tests 1 and 2 and a test proposed by Clayton and described by Campbell (1963, p. 225) were performed on a smaller sample of 100 sets of generated time series.][1]

For each test of significance, the percent of t's or F's which exceeded the theoretical 1% and 5% values was determined. The complete sequence of operations was performed internally on the IBM 709 computer system programmed to perform the necessary operations. The basic steps in the computation procedures are summarized below:

(1) The "true" line pre- and post-X occasion values were determined on the basis of desired slope. Normal random errors yielding specified autocorrelation were added to the "true" line points to form the set of time series data. (A binary subprogram from the Vogelback computing center (NU-0044) was used for the generation of normal random numbers. The mean of the Gaussian distribution approximates zero; the standard deviation was specified as described in (d) above). The program generated 1000 sets of time series for each type of error fluctuation specified.

(2) F and t values for the Mood, Walker-Lev 3, and Double Extrapolation tests of significance and serial correlation coefficients r(1), r(2),

r(3), and r(4) were computed for each of the 1000 sets of generated time series data. A data plot of the time series could also be obtained for each set.

(3) The 1000 t and F values for the Mood, Walker-Lev 3, and Double Extrapolation tests were sorted on the basis of magnitude. The ordered t and F values and/or the percents of t's and F's above the tabled critical values were printed out. The correlations between the t and F values and the autocorrelation coefficients were determined.

RESULTS

The results of the present study indicated that the three main tests of significance (Mood test, Walker-Lev Test 3, and Double Extrapolation Technique) are appropriate for use on data of interrupted time series form. However, the results also indicated that when statistical dependency between measurements exists use of these tests with the tabled critical values will yield significant results by chance alone more than the expected one percent and five percent of the time. This was particularly true for the Walker-Lev Test 3 and the Double Extrapolation Technique.

These results are summarized in Table 1 where the percent of F's and t's above the tabled 1% and 5% critical values are given for the three statistical tests of significance. These percent values were obtained from the

Insert Table 1 about here

number of instances of significance in 1000 sets of generated Monte Carlo time series. Four degrees of dependency between points and total numbers of time series occasion values of 20, 40, and 200 are represented in Table 1 (for a total N of 200, only the independent error and the lag three correlated error Monte Carlo generations were available). As indicated in Table 1, when no significant autocorrelation exists between points, that is, the errors are independent, the alpha values are approximately the theoretical

Supplementary Note:

Since the research reported herein was done, a highly relevant test of significance has been reported:

G. E. P. Box and George C. Tiao, "A change in level of a non-stationary time series," Biometrika, 1965, 52, 181-192.

In particular, the Box and Tiao approach avoids the assumption of linearity, in exchange for other, probably more reasonable, assumptions. What is needed is a computer program for the exact distribution computation by numerical methods for their test when δ and γ_0 are unknown (pp. 189-191), plus a testing of the formula against null models such as those used here, plus a testing of our linear formulas against the null data generated according to the Box and Tiao assumptions.

Table 1

Percent of F's and t's Above One and Five Percent Critical Values as a Function of Degree of Correlated Error

Test of Significance	Total Occasions in Time Series	One Percent				Five Percent			
		Independent Error	Correlated Error			Independent Error	Correlated Error		
			Lag 1	Lag 2	Lag 3		Lag 1	Lag 2	Lag 3
Mood Test	N = 20	1.10*	1.20	2.00	2.10	5.30	5.40	7.00	9.50
	N = 40	1.00	1.50	1.90	1.70	5.80	5.50	5.90	9.10
	N = 200	.90	--	--	.90	5.20	--	--	5.80
Walker-Lev Test 3	N = 20	1.10	2.80	5.90	9.20	5.60	9.60	17.30	19.80
	N = 40	1.40	3.30	8.50	13.20	5.20	11.50	20.20	25.30
	N = 200	1.40	--	--	17.50	5.90	--	--	28.80
Double Extrapolation	N = 20	1.10	2.90	7.40	11.10	5.30	11.10	19.20	23.40
	N = 40	1.50	3.60	9.50	15.10	5.00	11.60	20.60	27.10
	N = 200	1.40	--	--	17.70	5.70	--	--	28.40

*Each percent based on number of significant instances in 1000 sets of generated time series of true line slope = 1.00 and total error variance = 5.00.

values expected. When the assumption of independence is not met (correlated error), the percent of significant time series exceeds the expected one percent and five percent values. The percent of significant instances increases with increasing autocorrelation between points, particularly for the Walker-Lev Test 3 and Double Extrapolation Technique.

The slope of the true line and the total error variance about the true line had no effect on the percent of F's and t's above the tabled critical values. The percents of false-positives were similar for true line slopes of 0, 1, and 20 and for total error variances of 1.00 and 5.00. These results were obtained for all three tests of significance.

Although the true line slope and total error variance had no effect on the number of false-positives, the total number of occasions in the time series did produce an effect. As can be observed in Table 1, the percent of significant instances tends to vary with the total N of the series. This is particularly evident for the Walker-Lev Test 3 and Double Extrapolation Technique performed on time series with correlated error for three lags. The greater error in regression estimates with smaller N's is most likely the crucial factor in this effect.

The effect of the total number of occasion values in the time series was also evident in departures of the obtained average autocorrelation coefficients from their expected values. The expected values ρ_a of r_a is given by $\rho_a = cov\{Y_i Y_{i+a}\} / \sqrt{\sigma^2_{Y_i} \sigma^2_{Y_{i+a}}}$ and the standard error of r_a, by $\sigma_{r_a} = 1-\rho_a^2/\sqrt{N_s-1}$ where N_s refers to the sample size of 1000 in the present study. When expected values were compared with the obtained averages (over 1000 sets of generated time series) the obtained averages were less than the expected values for the smaller N's of 20 and 40. The obtained averages

approached the expected values for N of 200. Table 2 indicates this relationship between expected value and obtained averages for the first autocorrelation coefficient r_1. Similar results were obtained for the autocorrelation coefficients r_2, r_3, r_4.

Insert Table 2 about here

On the basis of the Monte Carlo findings of elevated alpha values for time series exhibiting proximally correlated error, we present in Table 3 new critical values for the three tests of significance. These new critical values approximate more closely the expected one and five percent significance levels. The new critical values were obtained by the method of basic linear interpolation from the IBM printout giving the "Percent of $\underline{z}$'s and $\underline{t}$'s in intervals of specific widths" for each test of significance. The values are based on Monte Carlo generations of 1000 sets of time series for each type of error fluctuation. As shown in Table 3, when time series

Insert Table 3 about here

observations are not correlated (assumption of independence met) the critical values which yield significant results one and five percent of the time by chance alone are, in general, similar to the tabled critical values for all three statistical tests of significance. However, when a significant autocorrelation exists between points, that is, when a given observation is dependent upon preceding observations, the critical values yielding significant results one and five percent of the time are similar to the tabled values for the Mood test only. For both the Walker-Lev Test 3 and Double Extrapolation Technique, the critical values increase with increasing autocorrelation.

The new critical values (Table 3) are plotted in Figures 1-6 as a

Table 2

Means and Standard Deviations of the First Autocorrelation Coefficient r_1

		Obtained Values			Expected Values
Error Type		N = 20	N = 40	N = 200	
Independent	$\bar{x}$	-.11	-.05	-.01	.00
	SD	.21	.16	.07	.03
Correlated Lag 1	$\bar{x}$	.18	.26	--	.33
	SD	.20	.13	--	.03
Correlated Lag 2	$\bar{x}$	.31	.41	--	.50
	SD	.22	.14	--	.02
Correlated Lag 3	$\bar{x}$	.37	.49	.58	.60
	SD	.24	.14	.06	.02

Table 3

Critical Values Yielding One and Five Percent Significance Levels

Test of Significance		One Percent			Five Percent		
		N = 20*	N = 40	N = 200	N = 20	N = 40	N = 200
Mood Test	Degrees of Freedom for t	df=8	df=18	df=98	df=8	df=18	df=98
	Tabled Critical Values of t	3.36	2.88	2.60	2.31	2.10	1.99
	Independent Error	3.49	2.94	2.59	2.36	2.23	2.02
	Correlated Error-Lag 1	3.32	3.08	----	2.35	2.18	----
	Correlated Error-Lag 2	3.74	3.28	----	2.51	2.22	----
	Correlated Error-Lag 3	3.62	3.30	2.59	2.77	2.42	2.08
Walker-Lev Test 3	Degrees of Freedom for F	df=1,17	df=1,37	df=1,197	df=1,17	df=1,37	df=1,197
	Tabled Critical Values of F	8.40	7.35	6.76	4.45	4.10	3.89
	Independent Error	8.58	8.58	8.25	4.72	4.20	4.29
	Correlated Error-Lag 1	10.68	11.06	----	6.22	5.96	----
	Correlated Error-Lag 2	17.25	18.12	----	9.10	9.25	----
	Correlated Error-Lag 3	23.75	24.75	24.75	11.25	13.44	14.04
Double Extrapolation	Degrees of Freedom for t	df=16	df=36	df=196	df=16	df=36	df=196
	Tabled Critical Values of t	2.92	2.70	2.58	2.12	2.03	1.96
	Independent Error	3.03	2.90	2.85	2.14	2.04	2.05
	Correlated Error-Lag 1	3.52	3.36	----	2.56	2.49	----
	Correlated Error-Lag 2	4.64	4.46	----	3.24	3.16	----
	Correlated Error-Lag 3	5.14	4.98	4.95	3.57	3.75	3.73

*N refers to the total number of occasions in the generated time series. Degrees of freedom and critical values for the Mood test are based on 10, 20, and 100 pre-change points. The total series points of 20, 40, and 200 are used for the Walker-Lev Test 3 and Double Extrapolation Technique.

function of the obtained autocorrelation coefficient. As discussed above, the obtained average r_1 is less than the expected value due to greater error in regression estimates with small N. Figures 1-6 may be used to obtain approximate one percent and five percent critical values when correlated error exists between points. The new critical values are found by interpolating on the abscissa using the obtained first autocorrelation coefficient and on the face of the graph using N. For the Mood test N refers to the number of pre-change points; for the Walker-Lev Test 3 and Double Extrapolation Technique N refers to the total number of occasions in the time series. The new critical value is read from the ordinate.

Insert Figures 1-6 about here

Some Utilizations

The tests of significance which we are offering are being applied in situations in which the visual impression is the usual basis of interpretation. In some sense, the tests are designed to imitate such judgments, making more precise and rationalizing the criteria involved. Thus, the more variable the pre-change points, the larger the cross-treatment change must be to "appear" significant. And holding this constant at a small variability, the longer the sampling of pre- and post-change observations, the more confident we are that a given trans-treatment shift is truly exceptional, is more than a coincidence. These ingredients feature prominently in the tests we have examined.

In the present study, we have examined only linear hypotheses. Many times a significant effect in terms of linear hypotheses will appear upon graphic inspection to be a homogenous curvilinear process with no discontinuity

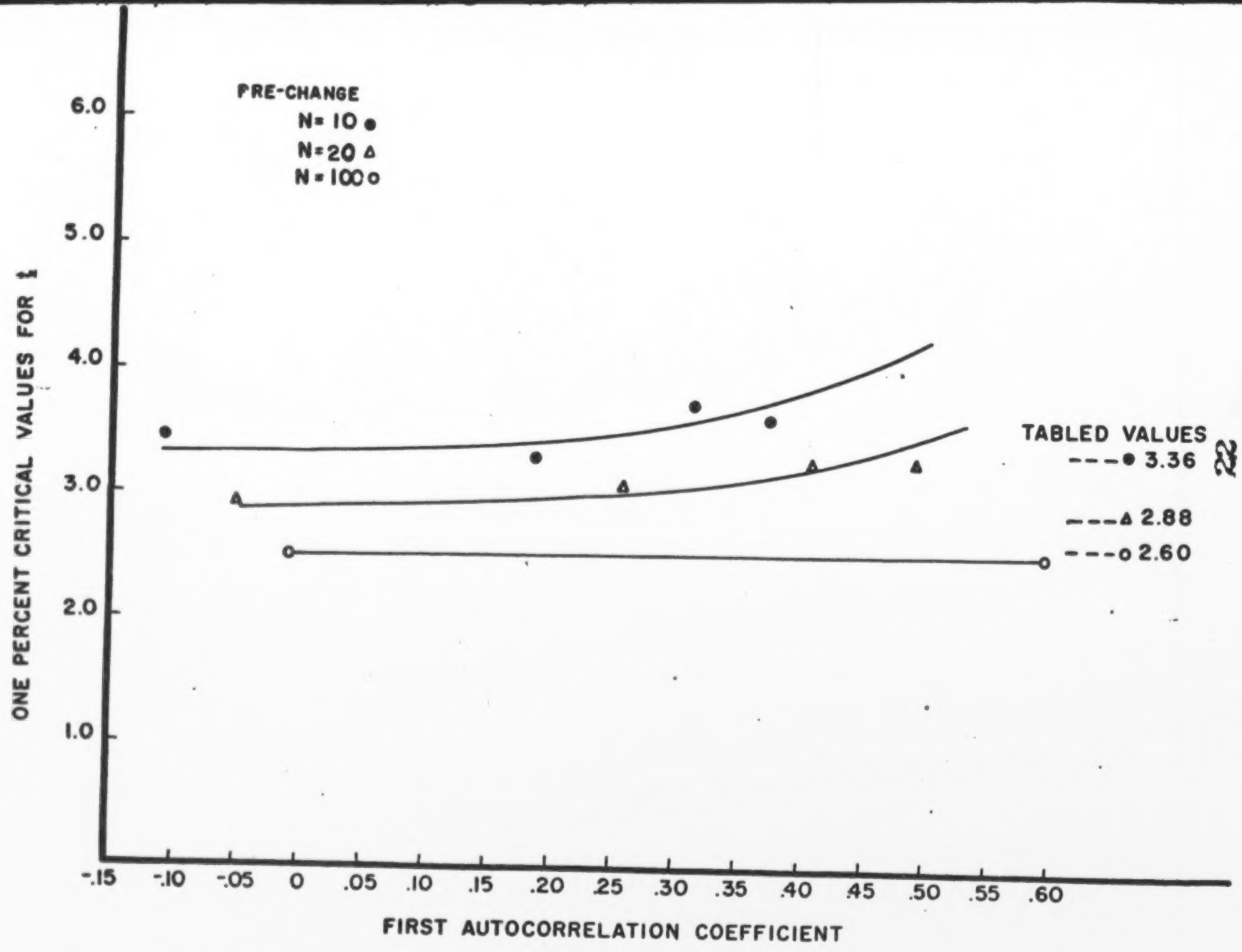

Fig. 1. Mood test: Obtained one percent critical values of t as a function of the first autocorrelation coefficient.

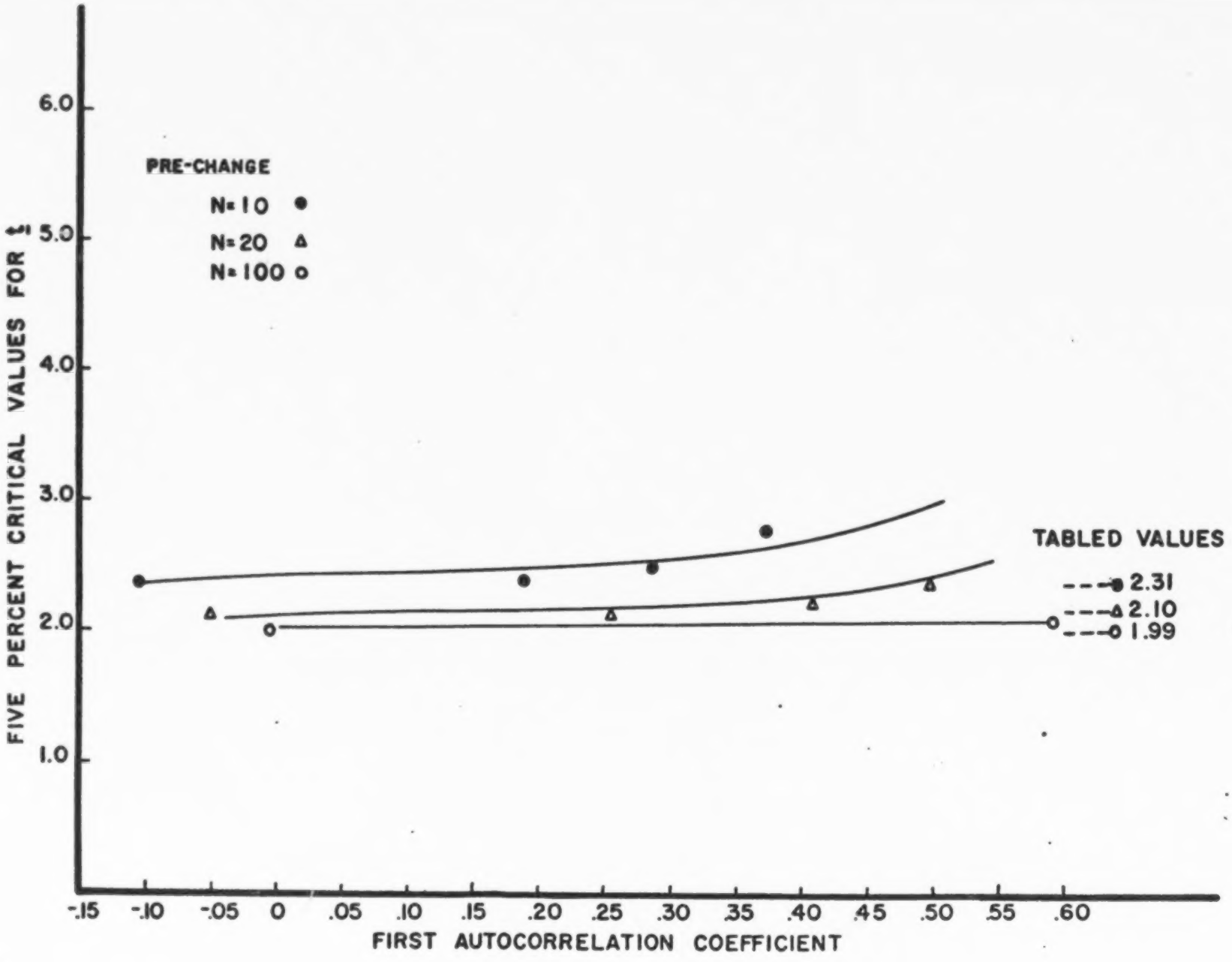

Fig. 2. Mood test: Obtained five percent critical values of t as a function of the first autocorrelation coefficient.

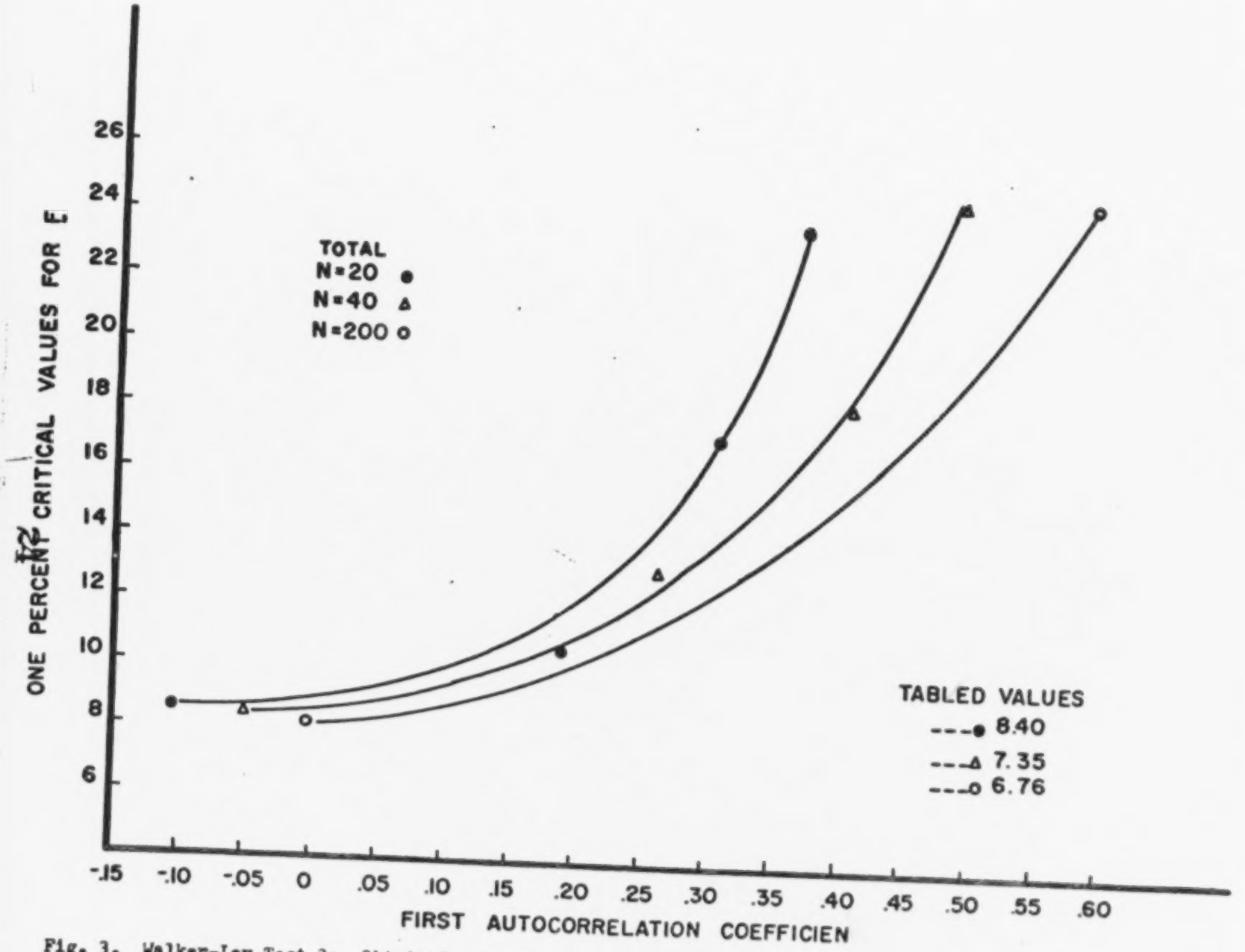

Fig. 3. Walker-Lev Test 3: Obtained one percent critical values of F as a function of the first autocorrelation coefficient.

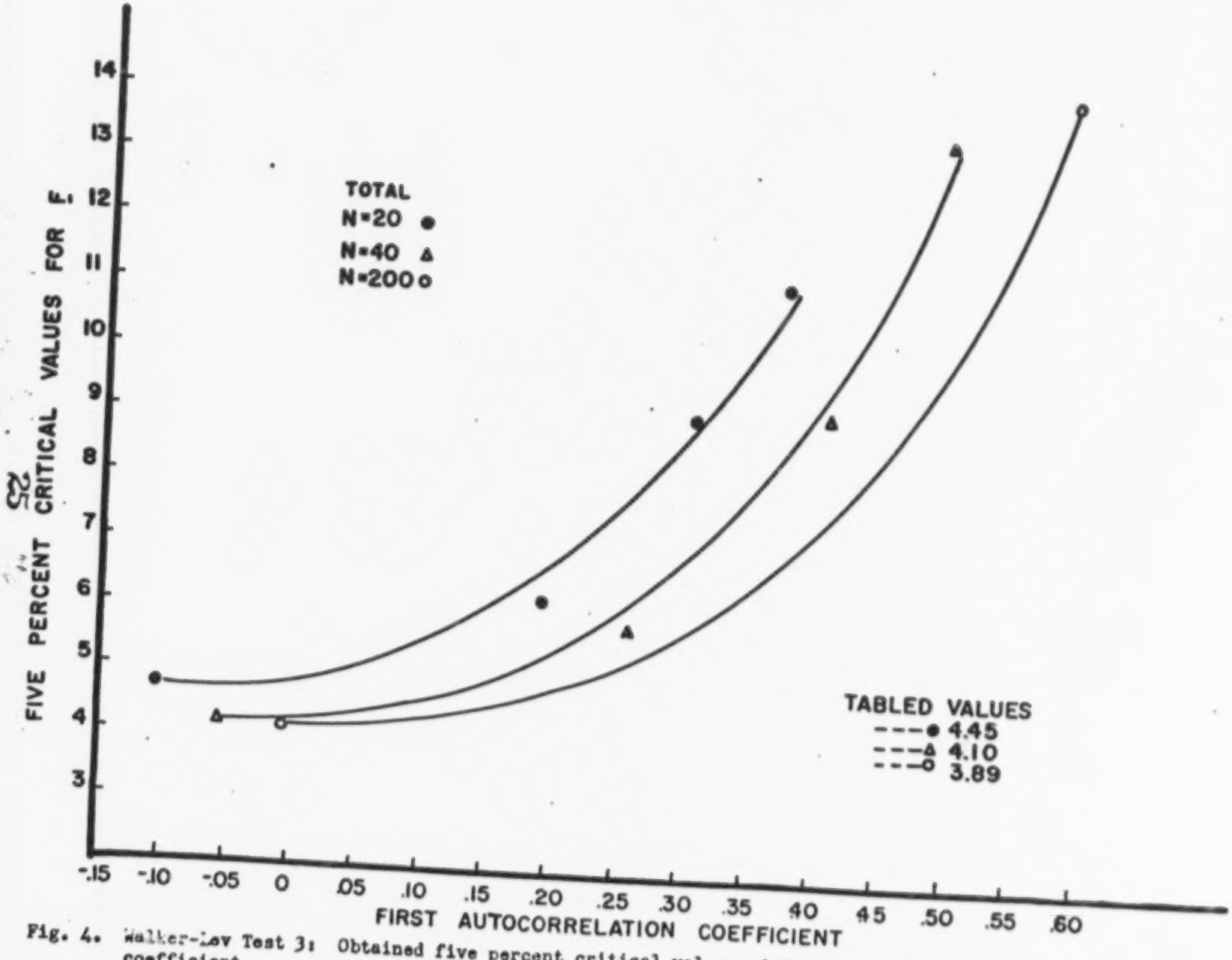

Fig. 4. Walker-Lev Test 3: Obtained five percent critical values of F as a function of the first autocorrelation coefficient.

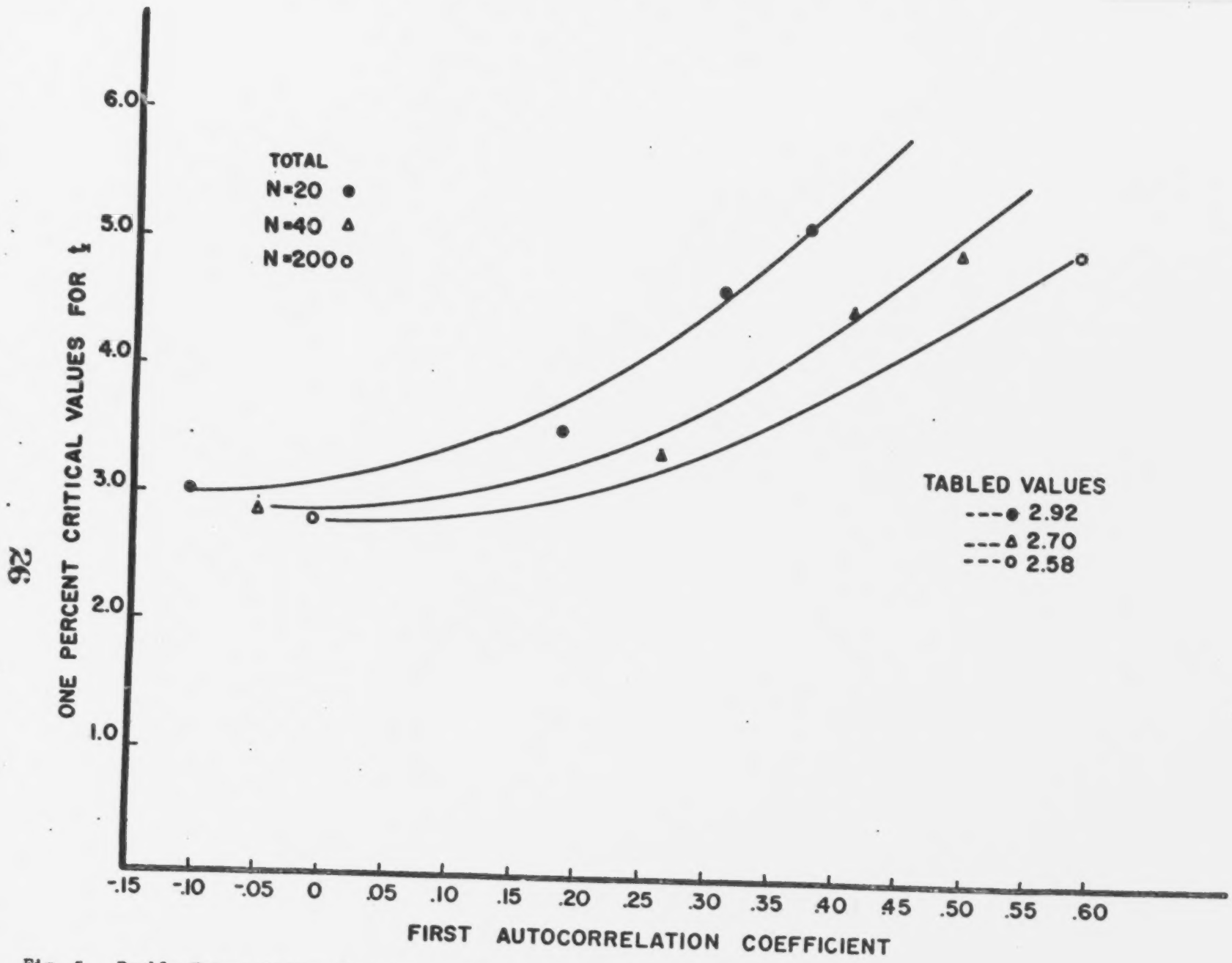

Fig. 5. Double Extrapolation Technique: Obtained one percent critical values of t as a function of the first autocorrelation coefficient.

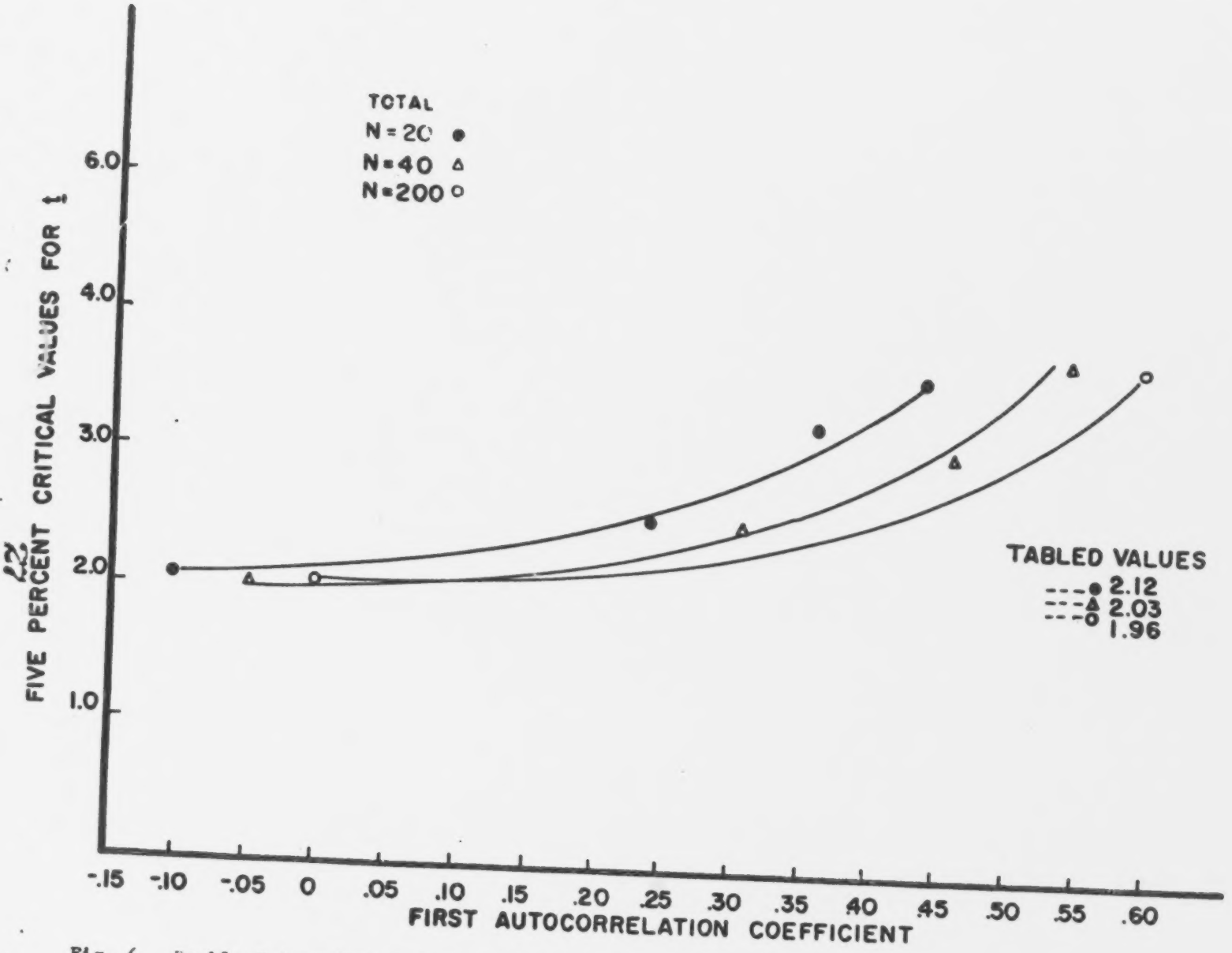

Fig. 6. Double Extrapolation Technique: Obtained five percent critical values of t as a function of the first autocorrelation coefficient.

at the treatment point. Usually one will not have sufficient degrees of freedom to test curvilinear hypotheses. For these and other reasons, it seems best to accompany tests of significance with graphic presentations.

In Figures 7, 8, 9, and 10 we present "significant" time series generated by the Monte Carlo of null conditions, one each for independent error and correlated error of one, two, and three lags. These graphs give some indication of the effect of proximal autocorrelations. In all four figures, the true line slope is 1.0, the total error variance about the true line is 5.00, and the total number of occasions is 40 (20 pre-change and 20 post-change). The specific data on the tests of significance is as follows:

Insert Figures 7, 8, 9, and 10 about here

For Figure 7 (independent error) the t value for the Mood Test was 2.55, the F value for the Walker-Lev Test 3 was 6.11, and the t value for the Double Extrapolation Technique was 2.45. The values of the autocorrelation coefficients based on departures from the line of best fit for the total series were $r_1 = .07$, $r_2 = -.24$, $r_3 = -.08$, $r_8 = -.07$. In Figure 8 (correlated error of lag 1) the t value for the Mood test was 2.34, the F value for the Walker-Lev Test 3 was 6.13, and the t value for the Double Extrapolation Technique was 2.45. The values of the autocorrelation coefficients were $r_1 = .45$, $r_2 = -.22$, $r_3 = -.44$, $r_4 = -.26$. In Figure 9 (correlated error of lag 2) the t value for the Mood test was 3.86, the F value for the Walker-Lev Test 3 was 4.66, and the t value for the Double Extrapolation Technique was 2.16. The values of the autocorrelation coefficients were $r = .51$, $r = .25$, $r = -.06$, $r = -.15$. In Figure 10(correlated error of lag 3) the t value for the Mood test was 3.54, the F value for the Walker-Lev Test 3 was 7.45, and the t value for the Double Extrapolation Technique was 2.88.

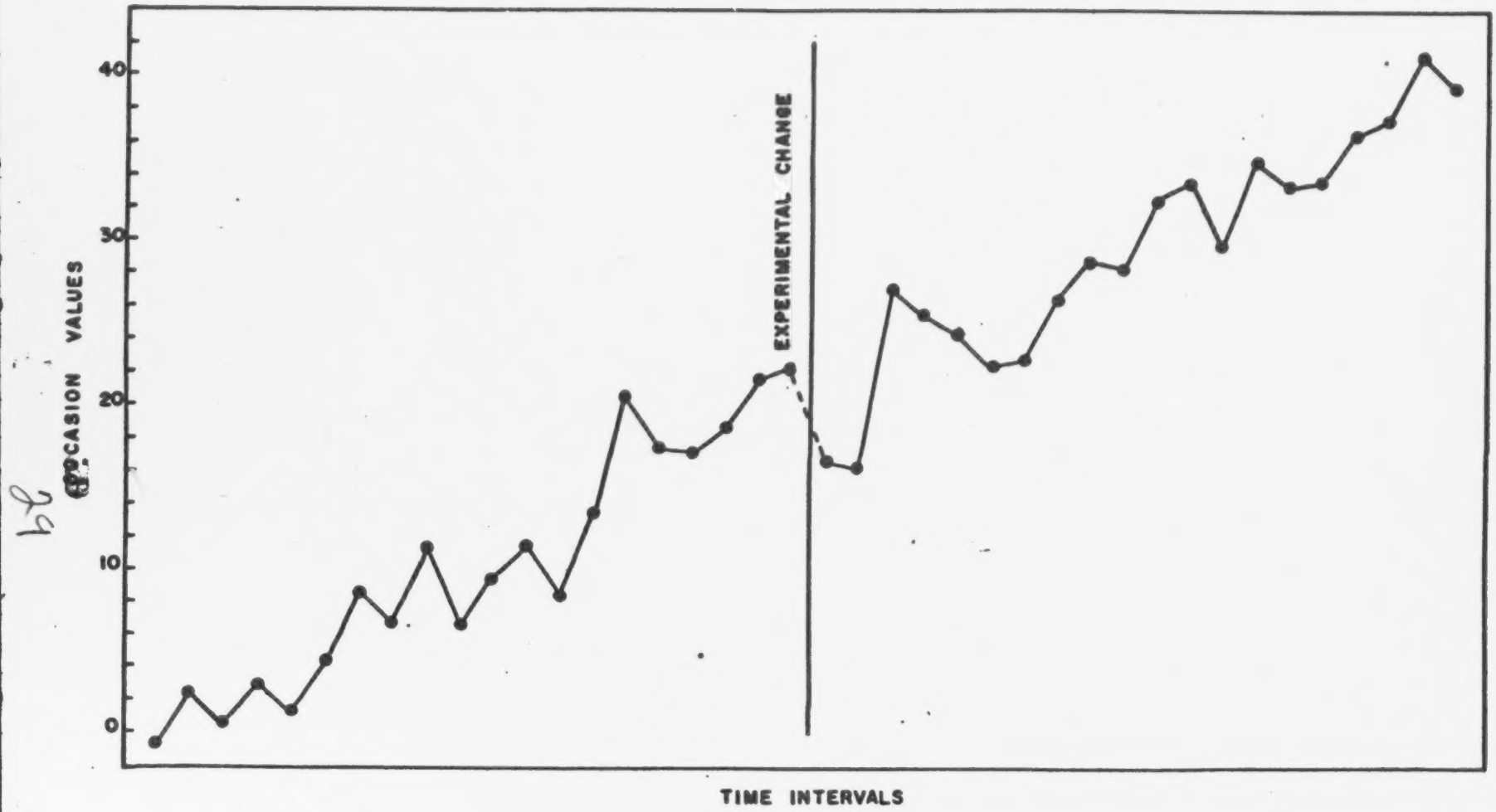

Fig. 7. A significant Monte Carlo generated time series for independent error.

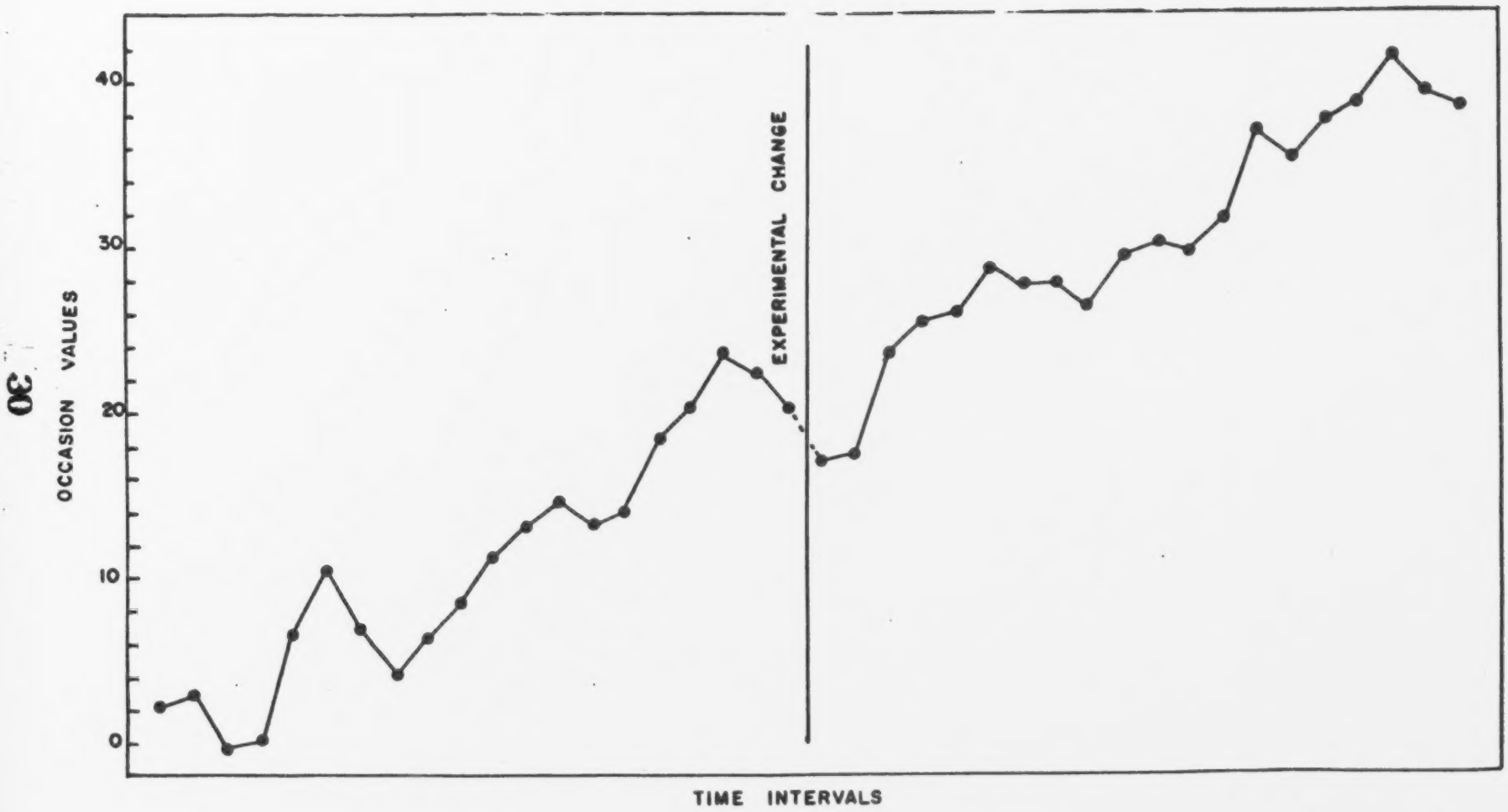

Fig. 8. A significant Monte Carlo generated time series for correlated error of one lag.

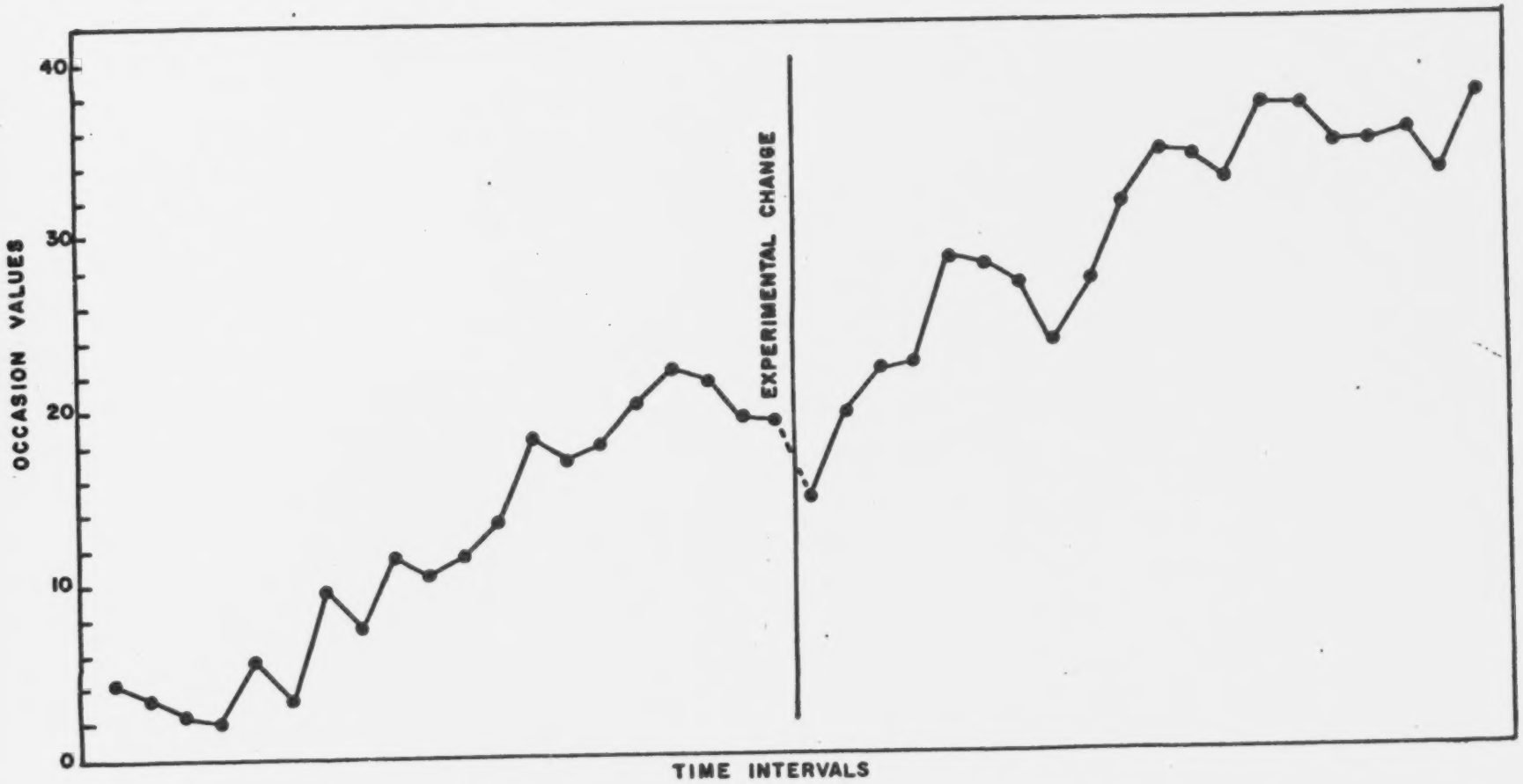

Fig. 9. A significant Monte Carlo generated time series for correlated error of two lags.

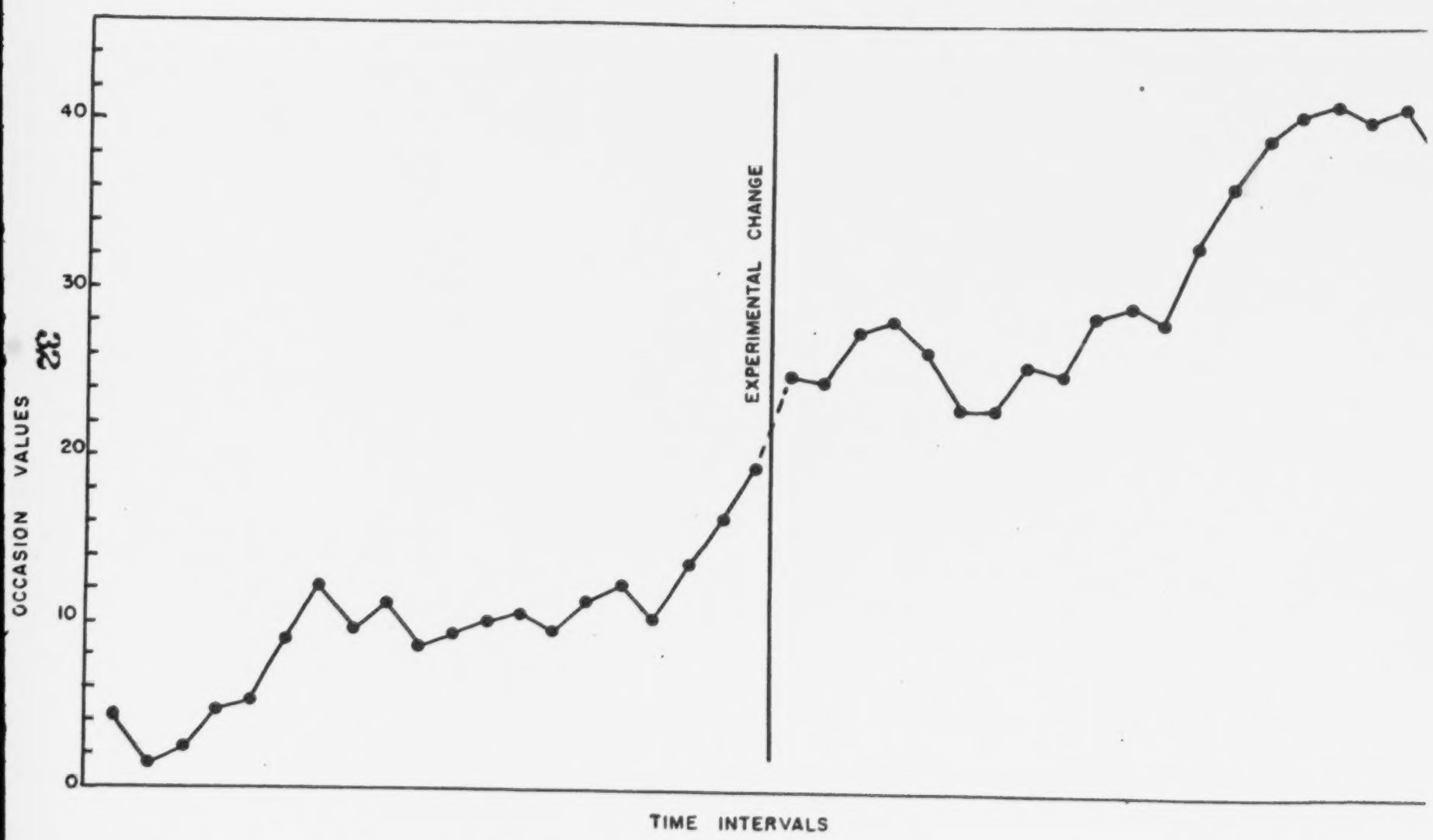

Fig. 10. A significant Monte Carlo generated time series for correlated error of three lags.

The values of the autocorrelation coefficients were $r_1 = .76$, $r_2 = .50$, $r_3 = .20$, $r_4 = -.14$.

To further illustrate utilizations of these tests of significance, in Figure 11, 12, 13, and 14 we present actual data. In general, these figures represent instances in which the tests of significance confirm judgments of effect which had been made on the basis of visual inspection alone.

Insert Figures 11, 12, 13, and 14 about here

Figures 11 and 12 represent Chicago crime rate statistics for the categories of "Larceny under $50" and "Murder and non-negligent manslaughter," respectively (from Uniform Crime Reports for the U.S., 1942-1962). On February 22, 1960, it was announced that Orlando Wilson had been selected as Superintendent of the Chicago Police Force (New York Times, February 23, 1960, page 1). He became acting commissioner on March 2, 1960. It was subsequently reported that "Recorded crime in Chicago increased 83.7% in the first 10 months of 1960, police statistics indicated today. The officials refused to say how much of the increase was due to a higher crime rate and how much to improved record keeping by police" (New York Times, December 17, 1960, page 34). In Figure 11, all statistical tests indicate that the observed increase in recorded "Larceny under $50" was significant. The *t* value for the Mood test was 12.54, the *F* value for the Walker-Lev Test 1 was 51.59, the *F* value for the Walker-Lev Test 3 was 157.76, and the *t* value for the Double Extrapolation Technique was 9.12. The value of the first autocorrelation coefficient based on departures from separate pre- and post-change regression lines was .43. However, in Figure 12 where the effect seems to start several observations before Orlando Wilson's appointment, none of the statistical tests were significant. The *t* value for the Mood

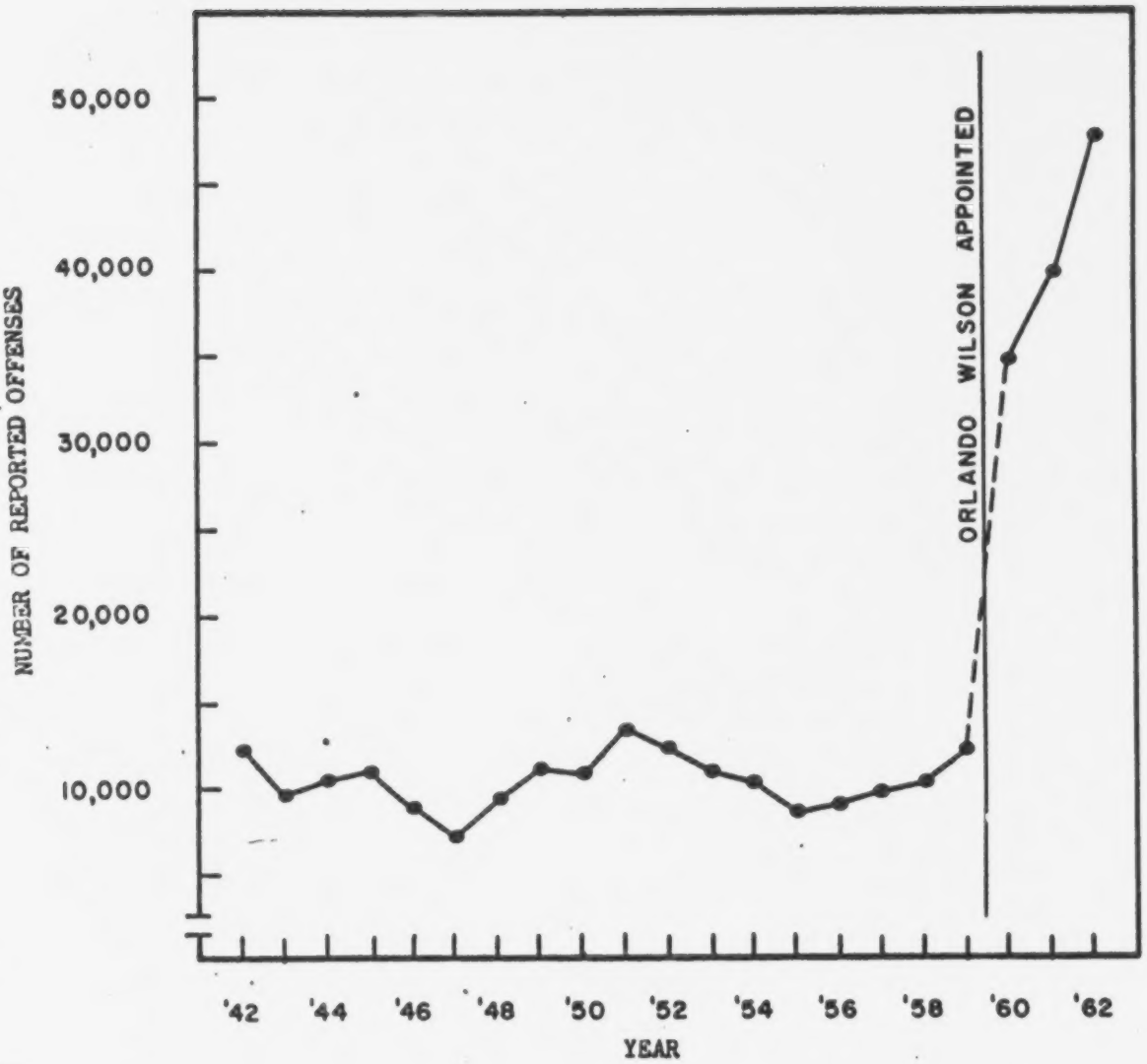

Fig. 11. Number of reported larcenies under $50. Chicago, Illinois. 1942-1962 (from Uniform Crime Reports for the United States, 1942-62).

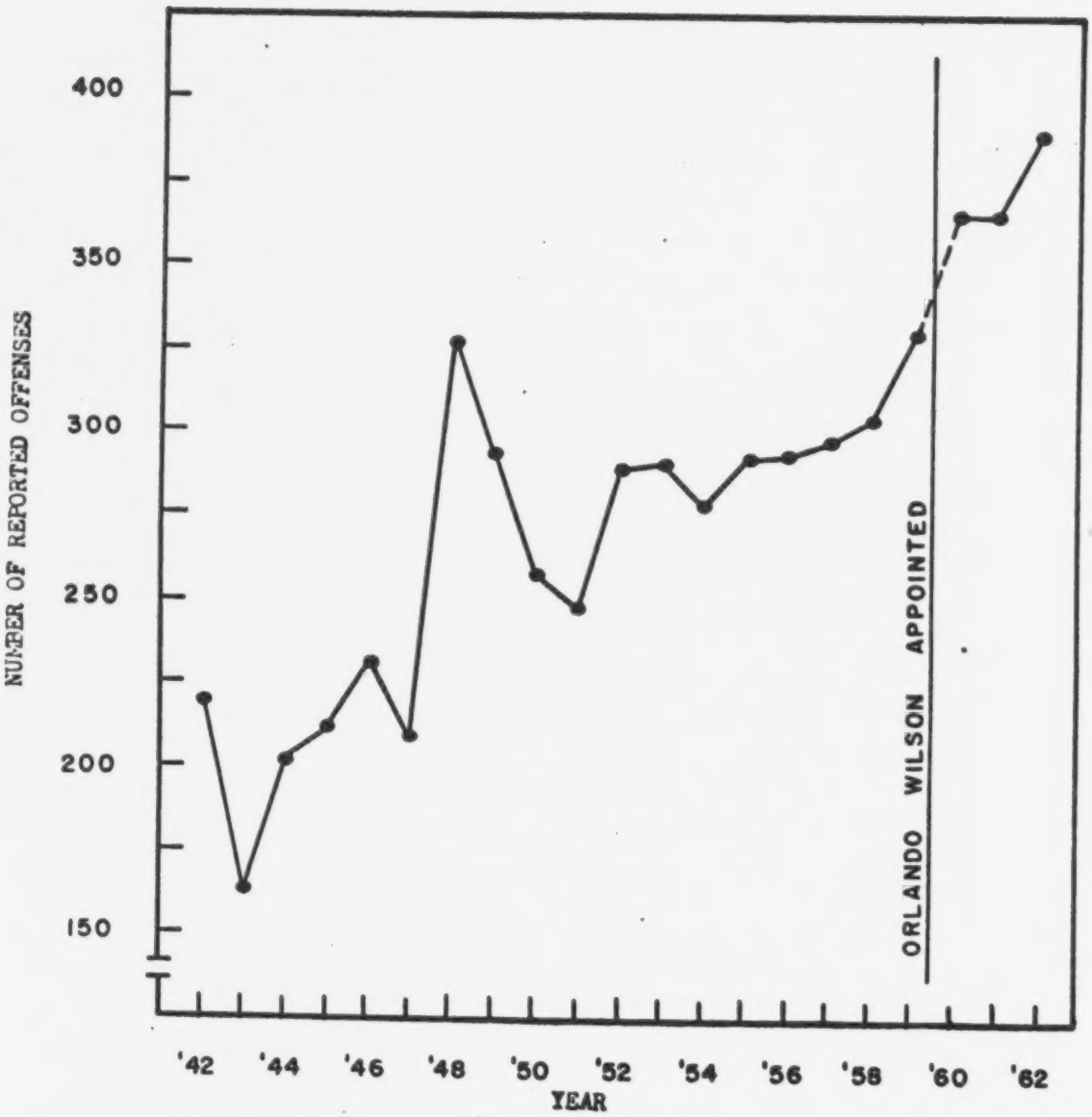

Fig. 12. Number of reported murders and non-negligent manslaughters. Chicago, Illinois. 1942-1962 (from Uniform Crime Reports for the United States, 1942-62).

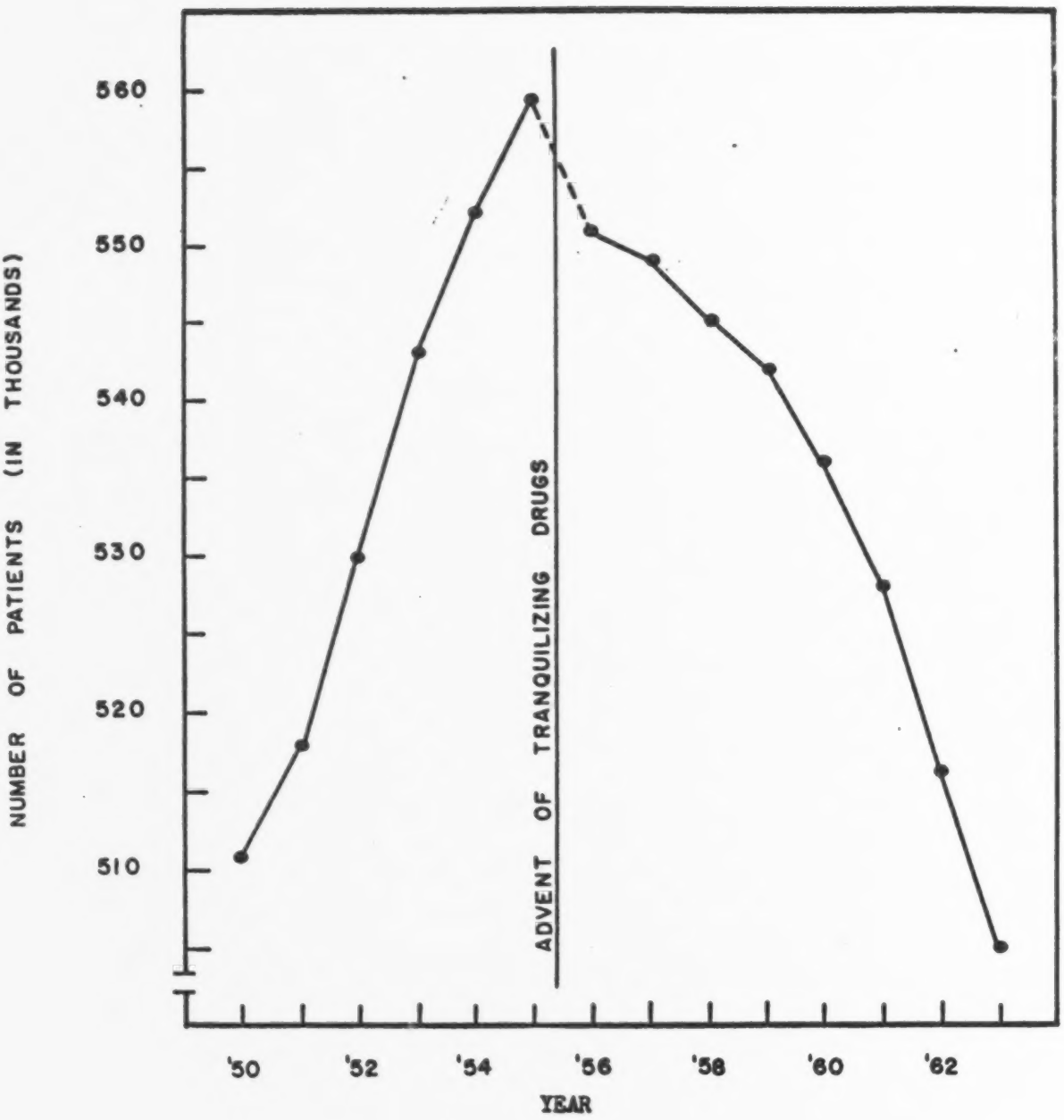

Fig. 13. Number of hospitalized mental patients in the United States before and after the advent of tranquilizing drugs (from Britannica Book of the Year, 1965).

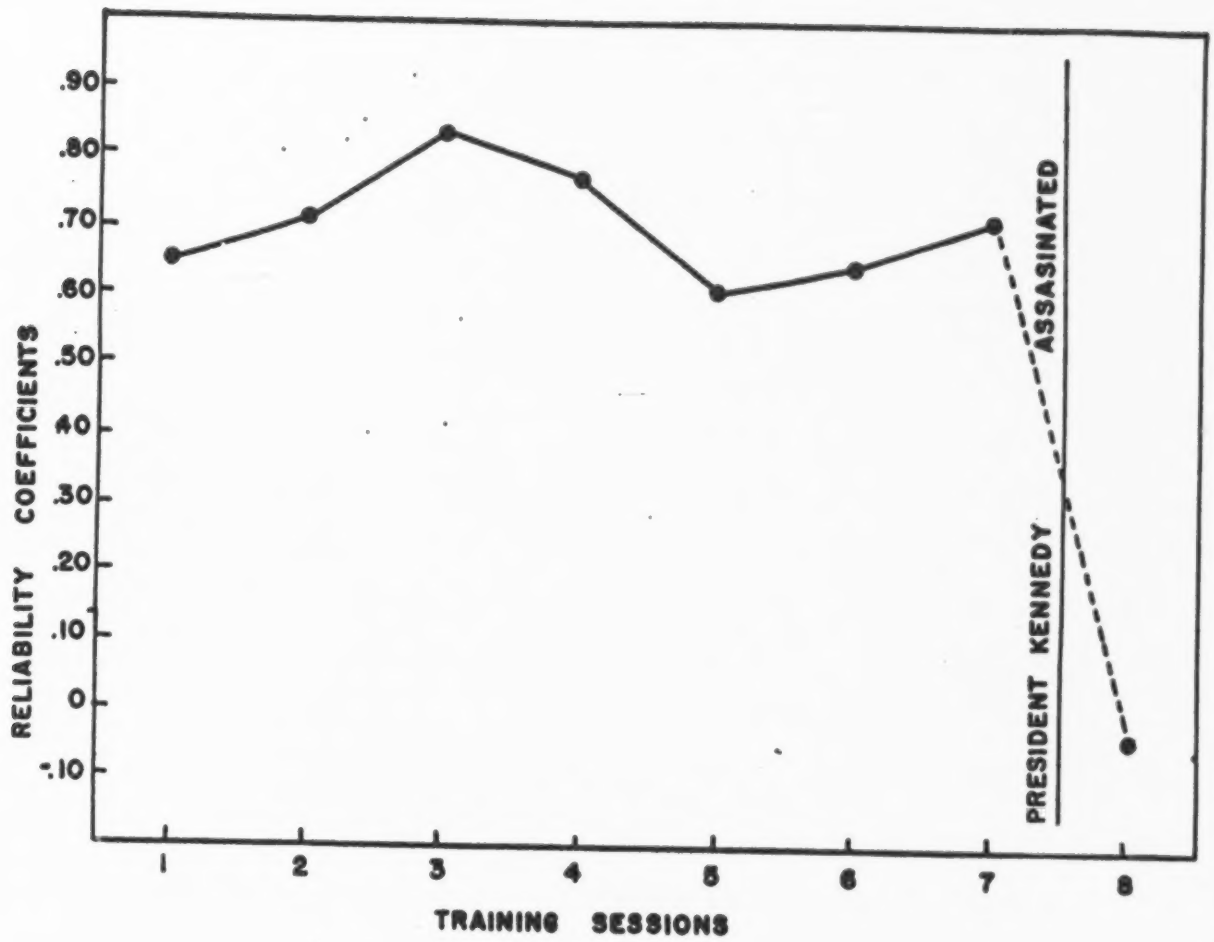

Fig. 14. Reliability coefficients for the structure dimension of the Leadership Opinion Questionnaire for training sessions before and after the assasination of President Kennedy (from Ayers, 1964).

test was 1.10, the F values for the Walker-Lev tests 1 and 3 were .03 and 2.69, respectively, and the t value for the Double Extrapolation Technique was .82. The value of the first autocorrelation coefficient based on departures from the total series regression line was .15.

Figure 13 represents the number of hospitalized mental patients in the United States before and after the advent of the use of tranquilizing drugs (from Britannica Book of the Year, 1965). The significant statistical tests were the Walker-Lev Test 2 (F = 213.75) and the Mood test (t = 7.40). The F value for the Walker-Lev test 3 was .26 and the t value for the Double Extrapolation technique was 1.38. The value of the first autocorrelation coefficient based on differences from the separate pre- and post-change regression lines was .52.

Figure 14 represents reliability coefficients for the Structure Dimension of a Leadership Opinion Questionnaire which was administered at the beginning and end of each of eight week-long training sessions. The training session groups ranged in size from 22 to 55 participants. Classes for the week-long sessions were given in the Spring and Autumn months of 1963 and are indicated in chronological order on the abscissa of the graph. For the eighth training session, pretesting took place on November 18, 1963, whereas posttesting occurred on Friday afternoon, November 22, at the close of the session. During the lunch period, 1:30-2:30 P.M., the participants had watched and listened to the memorable events being reported from Dallas, Texas (from Ayers, 1964). Since only one post-change point is given, only the Mood test of significance is applicable. The obtained t value for the Mood test was 6.24 with 5 df, thus, confirming the impression of effect. The value of the first autocorrelation coefficient based on differences from the

separate pre- and post-change regression lines was .12.

As the above examples of actual time series illustrate, the tests of significance are not equally applicable to all time series data. The several possible time series of Figure 15 indicate this. The time series of A1 and

Insert Figure 15 about here

A2 are instances of sustained post-change effect involving a change in intercept; whereas, B1 and B2 show an initial jump and subsequent return to pre-change conditions. If theoretical expectations are appropriate to using a series of points in the post-change period (as they would be in A1 and A2) the Walker-Lev Test 3 and Double Extrapolation Technique would be most suitable. These two tests would not be used in instances of theoretical expectations similar to B1 and B2 where only the Mood Test of the significance of the first-post change observation from a trend extrapolated from the pre-change observations would be appropriate.

A times series analysis seems most suitable in instances discussed above. In C1, C2, and C3 where the post-treatment effect involves a change in slope and, generally, of intercept, the possibility of a curvilinear relationship or cyclical trend is more difficult to rule out. A large number of pre- and post-treatment observations would be helpful in eliminating these possibilities. All three statistical tests and, in particular, the Walker-Lev Test 1, may be used, but they may yield conflicting results. For example, the double extrapolation technique can be used to indicate if the pre- and post-regression lines coincide at time t_o midway between the last pre-change and first post-change point. However, extension of the regression lines of C2 will show that the two lines coincide at time t_o, although pre- and post-regression lines are dissimilar.

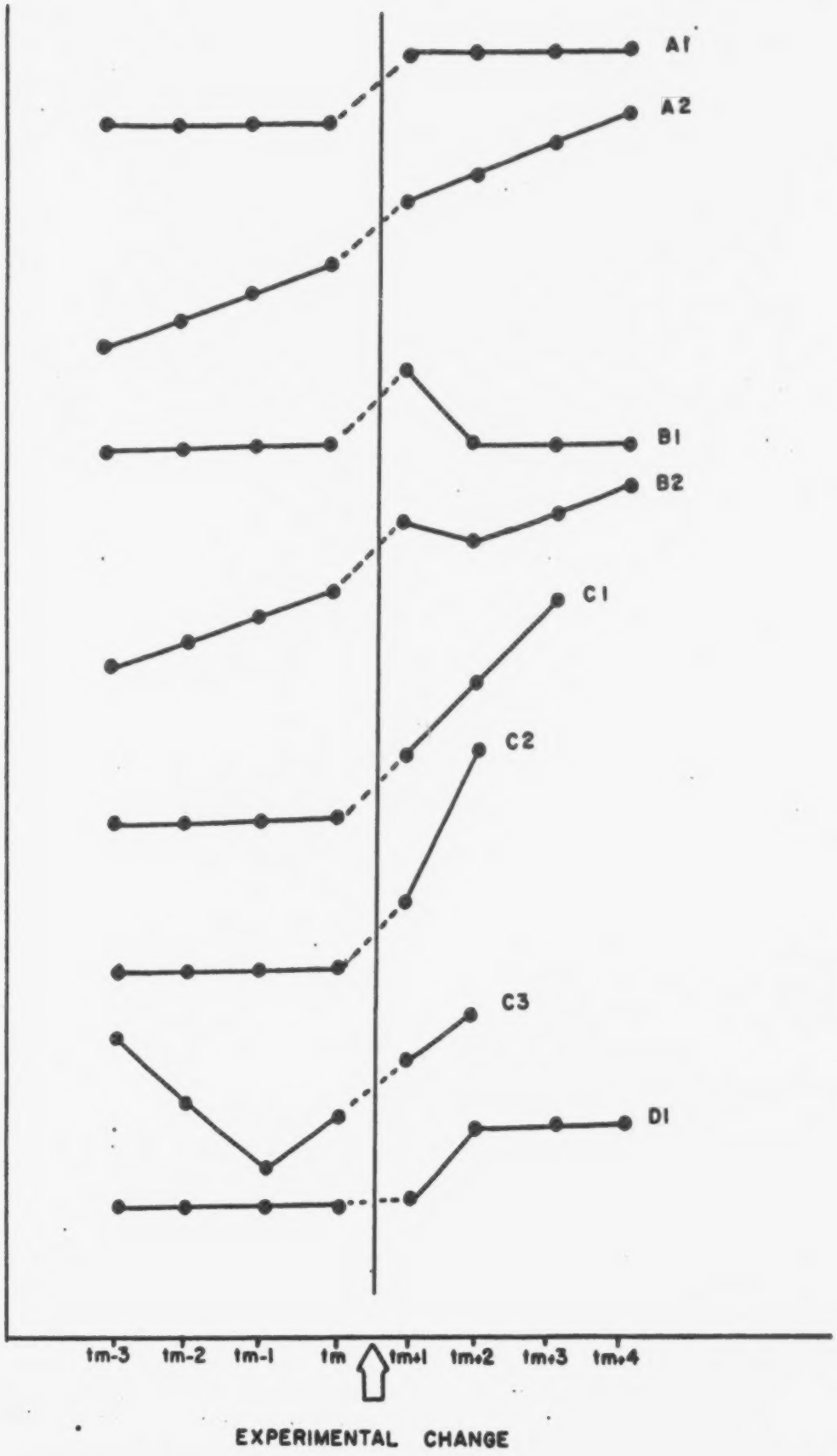

Fig. 15. Some possible outcome patterns for the interrupted time series experiment.

Similarly, the first post-change point may not differ greatly from that predicted by the pre-change values, yet a continuing and substantial increase or decrease in subsequent post-change values may suggest an effect.

In instances of delayed effect, such as D1, ambiguity is introduced into the interpretation of significance (as the time interval between the treatment and its effect increases the plausibility of rival hypotheses also increases). However, if the experimenter specifies in advance the exact relationship between the introduction of the treatment and the manifestation of its effect, the pattern indicated by time-series D1 could be almost as definitive as that in which immediate effect is expected.

References

Ayers, Arthur (Westinghouse Electric Corporation). Letter in American Psychol., 19, 1964, p. 353.

Britannica Book of the Year, 1965. Chicago, Encyclopedia Brittanica, 1965, p. 626.

Campbell, D. T. From description to experimentation. In C. W. Harris (Ed.), Problems in measuring change. Madison, Wis.: University of Wisconsin Press, 1963, pp. 212-242.

Campbell, D. T., and Stanley, J. S. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963, pp. 171-246.

Holtzman, W. H. Statistical models for the study of change in the single case. In C. W. Harris (Ed.), Problems in measuring change. Madison, Wis.: University of Wisconsin Press, 1963, pp. 199-211.

Mood, A. F. Introduction to the theory of statistics. New York: McGraw-Hill, 1950.

Sween, Joyce A., and Campbell, D. T. The interrupted time series as quasi-experiment: three tests of significance. Northwestern University, 1965.

Uniform Crime Reports for the United States. Issued by the Federal Bureau of Investigation, United States Dept. of Justice, Washington, D.C.

Walker, Helen M., and Lev, J. Statistical inference. New York: Holt, 1953.

Footnotes

[1]A preliminary investigation of the Clayton Test using 100 sets of generated time series (total N = 40, true line slope - 1.00, total error variance = 5.00) yielded alpha values greater than the expected one and five percents when no dependency between points was built in (independent error). The percents of significant instances exceeding the five percent tabled critical value of F were 17%, 36%, 53%, and 61% for independent error, lag 1, lag 2, and lag 3 correlated errors, respectively. In lieu of this preliminary finding, the Clayton test was not considered a useful significance test in the interrupted time series situation and it was eliminated from further Monte Carlo investigation.

In a similar preliminary investigation, the Walker-Lev Test 1 yielded 4% (independent error), 16% (lag 1 error), 26% (lag 2 error), and 34% (lag 3 error) false-positive instances above the five percent tabled critical value of F. Although the Walker-Lev Test 1 yielded results similar to those obtained in a preliminary investigation of the Walker-Lev Test 3, it was eliminated in the final 1000 set investigation. Preference was given to the Walker-Lev Test 3 because of the more restricted usefulness of Test 1. Since the Walker-Lev Test 1 is a test of slope differences only it is more difficult to rule out rival curvilinear hypothesis in cases of significance.

The Walker-Lev Test 2 was not used because a test of the null hypothesis of zero slope is not of general interest as a test of significance in the time series situation.